УДК 519.7

А.С. Пузик, В.В Курасова, Г.Г. Четвериков
Харьковский национальный университет радиоэлектроники, Украина
пр. Науки, 14, г. Харьков, 61000

АСПЕКТЫ ОРГАНИЗАЦИИ МОДЕЛИ ДАННЫХ ЭЛЕКТРОННОГО ТРЁХЪЯЗЫЧНОГО СЛОВАРЯ

O.S Puzik, V.V Kurasova, G.G. Chetverikov
Kharkiv National University of Radio Electronics, Ukraine
Nauky av, 14, Kharkiv, 61000

ASPECTS OF DATA MODEL ORGANIZATION OF ELECTRONIC TRILINGUAL DICTIONARY

Статья посвящена описанию аспектов организации модели данных для программной системы русско-украинско-английского терминологического словаря. В статье описан подход к разбору входных данных. Проведен анализ ошибок возникших при обработке входных данных и методы их решения. Приведено сравнение подходов для хранения данных для трехъязычных словарей. Преимуществом предложенной системы является равноправие языков, при котором основным языком можно назначать любой из представленных.

Ключевые слова: модель данных, лексикография, алгебра конечных предикатов, office automation, парсинг.

The article is devoted to describing of aspects of data model organization for software system of electronic Russian-Ukrainian-English terminological dictionary. The article describes approach to handle input data. Issues appeared during parsing of input data have been analyzed and identified ways to resolve them. Different approaches to store dictionary data were compared. The advantage of the system is the equality of languages, so any of dictionary languages may be set as main.

Keywords: data model, lexicography, algebra of finite predicates, office automation, parsing.

Введение

Современный мир невозможно представить без компьютеризированных систем. Они проникли во все сферы жизни от систем управления производственными процессами, до систем «умных» домов, на них основываются и простые веб-сайты, и сложные информационные порталы. Развитие лексикографии также не стоит на месте, работа лексикографа переводится в электронную форму, на смену бумажным словарям приходят электронные, которые доступны в том числе и через интернет. Компьютеризация процессов сбора текстовой информации позволяет упрощать анализ и дальнейшую ее обработку. Поэтому разработка электронных словарей является необходимой базой для дальнейшего развития систем интеллектуальной обработки естественной речи.

В данной статье предлагается подход к построению модели данных для трехъязычного терминологического словаря по информатике и радиоэлектронике. Кроме того, в данной статье описаны проблемы, связанные с представлением, обработкой и хранением данных. Описанные трудности решаются с помощью средств теории лексикографических систем [1, 2] и алгебры конечных предикатов (АКП) [3, 4].

Постановка проблемы и цели исследования

Задача построения электронного словаря представляет собой трудоёмкую многоэтапную процедуру. Исходными данными являются отсканированные страницы печатных словарей, которые необходимо перевести в электронную форму. Для этого необходимо решить проблему корректуры отсканированных текстов. После получения сырых данных, возникают проблемы, связанные с их представлением, обработкой и хранением. В данной работе требуется обработать отсканированные тексты словарных

статей печатного словаря, используя различные приемы, и исследовать аспекты организации модели данных для трёхъязычного словаря. Данные проблемы необходимо решить для русско-украинского словаря по информатике и радиоэлектронике. После перевода в электронную форму, двуязычный русско-украинский словарь необходимо преобразовать в трехъязычный русско-украинско-английский словарь.

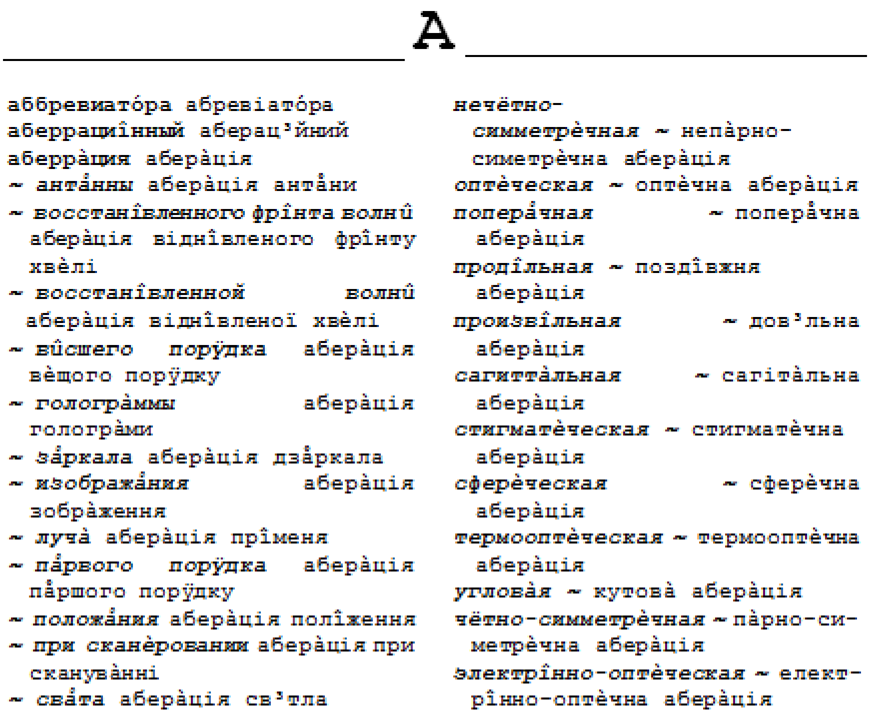
Подход к ообработке отсканированных текстов

Процесс создания электронных словарей обычно включает следующие этапы:
а) сканирование и распознавание словарных статей;
б) корректура полученного текста;
в) разбиение текста словаря на массив отдельных словарных статей;
г) декомпозиция массива словарных статей по формальным признакам [2].

Исходными данными словаря являются отсканированные и распознанные документы в формате «doc». Это - бинарный формат хранения файлов. используемый программой MSWord.

Принципом построения словаря является алфавитно-гнездовой [5]. Заголовочным словом является русское слово-термин. Гнездо включает терминологические словосочетания, элементом которых является заголовочный термин. Терминологические словосочетания строятся таким образом, чтобы тильда была на первом месте.

Напрямую эти данные из файла достать довольно затруднительно из-за внутренних особенностей формата «doc». Также сам распознанный текст находится в различных секциях документов и содержит большое количество ошибок распознавания, что затрудняет извлечение правильной информации (рис. 1).


А

аббревиатóра абревіатóра
аберрацiîнный аберац'йний
аберрàция аберàція
~ антåнны аберàція антåни
~ восстанîвленного фрîнта волнû аберàція віднîвленого фрîнту хвèлі
~ восстанîвленной волнû аберàція віднîвленої хвèлі
~ вûсшего порÿдка аберàція вèщого порÿдку
~ голограммы аберàція голограми
~ зåркала аберàція дзåркала
~ изображåния аберàція зображення
~ лучà аберàція прîменя
~ пåрвого порÿдка аберàція пåршого порÿдку
~ положåния аберàція полîження
~ при сканèровании аберàція при скануваннi
~ свåта аберàція св'тла
нечётно-симметрèчная ~ непàрно-симетрèчна аберàція
оптèческая ~ оптèчна аберàція
поперåчная ~ поперåчна аберàція
продîльная ~ поздîвжня аберàція
произвîльная ~ дов'льна аберàція
сагиттàльная ~ сагітàльна аберàція
стигматèческая ~ стигматèчна аберàція
сферèческая ~ сферèчна аберàція
термооптèческая ~ термооптèчна аберàція
угловàя ~ кутовà аберàція
чётно-симметрèчная ~ пàрно-симетрèчна аберàція
электрîнно-оптèческая ~ електрîнно-оптèчна аберàція

Рис. 1 Пример входных данных

Наименее трудоемким подходом для извлечения необходимой информации, в данном случае, является доступ к содержимому посредством технологии Office COM Automation [6]. MSWord предоставляет интерфейсы для доступа к содержимому документов. Таким образом, словарные статьи можно перевести в текстовый юникодный формат как промежуточный вариант. С одной стороны, данные получаются

не отягощенными информацией о форматировании, с другой стороны, юникод позволяет сохранять все символы любого языка, благодаря чему тексты можно использовать для дальнейшей обработки.

После перевода в текстовый формат, в словарных статьях обнаружилось много ошибок из-за неправильно распознанных символов, неверно распознанных знаков переносов, пустых строк и т.д. В то же время, в этих ошибках присутствовали определенные регулярности, что позволило организовать исправление данных ошибок в автоматизированном режиме. Каждый новый термин начинается с новой строки, а ошибочные символы «новой строки» можно выявить при помощи регулярных выражений. Все ударные буквы оказались неправильно распознанными, но единообразно, благодаря чему их удалось исправить простой заменой. Ударные буквы, после лексического разбора текста, помечаются символом «#».

При переносе данных из формата MSWord все словосочетания были пронумерованы, поэтому поиск неправильных вхождений переводов строк был возможен при помощи регулярных выражений типа «^\d\.$» («\d\.» – служебная цифра с точкой для нумерации перевода, добавленная в начало каждой строки). Аналогичным образом были выявлены неправильные скобки. В общем случае регулярные выражения неприменимы для определения скобок. Однако, такой подход оказался приемлем, поскольку отсутствовали сложные вложенные структуры скобок. Большинство неправильных скобок не имели пару. Обычно это случалось из-за того, что разрыв слова происходил в неправильном месте при распознавании.

Пример из разбитого на переводы строк, но необработанного, файла:

...

1.(моно, не)хроматѐческая □□(

2.моно, не)- хроматѐчна аберàція)

...

1.(-витîк

2.ампåр-витîк

...

В первом случае скобка из верхней строки должна принадлежать нижней, во втором скобка – артефакт распознавания. Такие случаи необходимо было найти и, соответственно, можно было использовать выражения типа «\([^\)]*?$».

Выявление неправильных дефисов оказалось самой трудоемкой частью. Чаще всего они приходились на перенос слова, из-за чего нарушалась структура изложения терминов. Благодаря этим нарушениям, при помощи регулярных выражений, такие места были найдены и исправлены почти для всех случаев. Такие проблемные символы были найдены при помощи регулярных выражений типа «-\s*$». Остальные дефисы пришлось исправлять в ручном режиме.

При дальнейшем анализе текстов, удалось выбрать отрасль, семантику (расширенное описание), изменяемую часть слов и прочее. В дальнейшем эти данные пригодились для заполнения внутренних структур словаря.

Модель данных трехъязычного словаря

После получения сырых распознанных и откорректированных данных, возникают проблемы, связанные с их обработкой, хранением и представлением пользователю. Одной из главных причин, возникающих на пути решения указанных проблем, является сложная структура лингвистического материала. Отсюда вытекает сложность организации модели данных. Часть проблем и подходов к их решению была описана для двуязычного словаря и лексикографических систем в целом [1, 2].

На верхнем уровне внутренняя структура двуязычного словаря обычно представляется следующим образом: в общих чертах, опуская особенности слов, каждому термину ставится в соответствие его переводной эквивалент или список эквивалентов, которые могут быть как синонимами, так и иметь другой смысл. Из набора таких эквивалентов состоит словарь. Обычно, один из языков является основным и переводы даются относительно этого языка. Данный случай – это пример отношения один ко многим. В случае, если перевод осуществляется в обе стороны, то появляется второй аналогичный список для другого языка. Для хранения содержимого словаря в электронной форме будут использованы таблицы для хранения переводных эквивалентов и связей между ними. При этом получаются связи многие ко многим. При переходе к созданию трёхъязычного словаря, особенно, если планируется, что все три языка должны быть равноправными, появляется проблема построения связей между переводными эквивалентами, поскольку количество связей растёт пропорционально количеству языков (рис. 2).

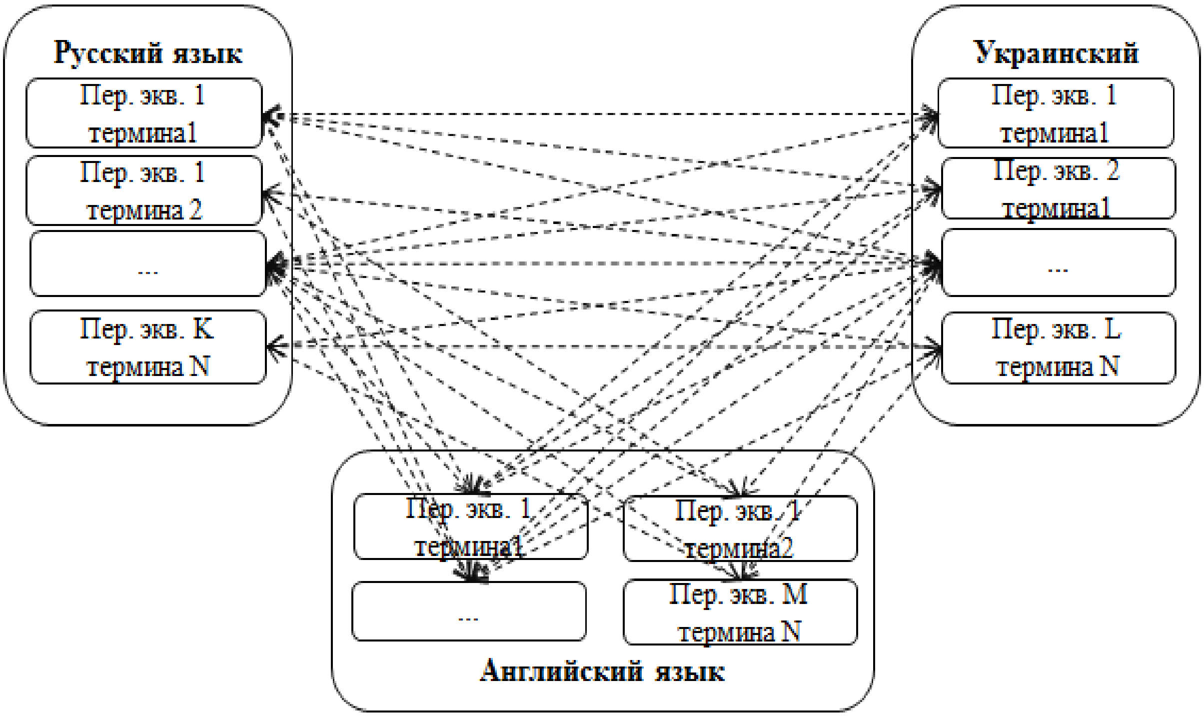


Рис. 2 Лавиноподобный рост количества связей при добавлении третьего языка

Одним из вариантов решения данной проблемы может являться вынос связей в отдельную сущность. Благодаря такому подходу, термины, терминологические словосочетания и связи между переводами для разных языков будут храниться отдельно, но упорядочено. При этом остается проблема семантического перевода для конкретного термина, то есть в некоторых ситуациях может потребоваться получить информацию только о какой-то конкретной семантике слова.

Другим подходом для решения данной проблемы предлагается введение дополнительного уровня косвенности, которым будет являться абстракция, обозначающая семантику термина. Это позволит, с одной стороны, перейти от отношения многие ко многим к отношению один ко многим (рис. 3), с другой стороны, появится четкая привязка термина к семантике, с третьей стороны, при дальнейшем развитии словаря, терминам можно будет легко добавлять толкования.

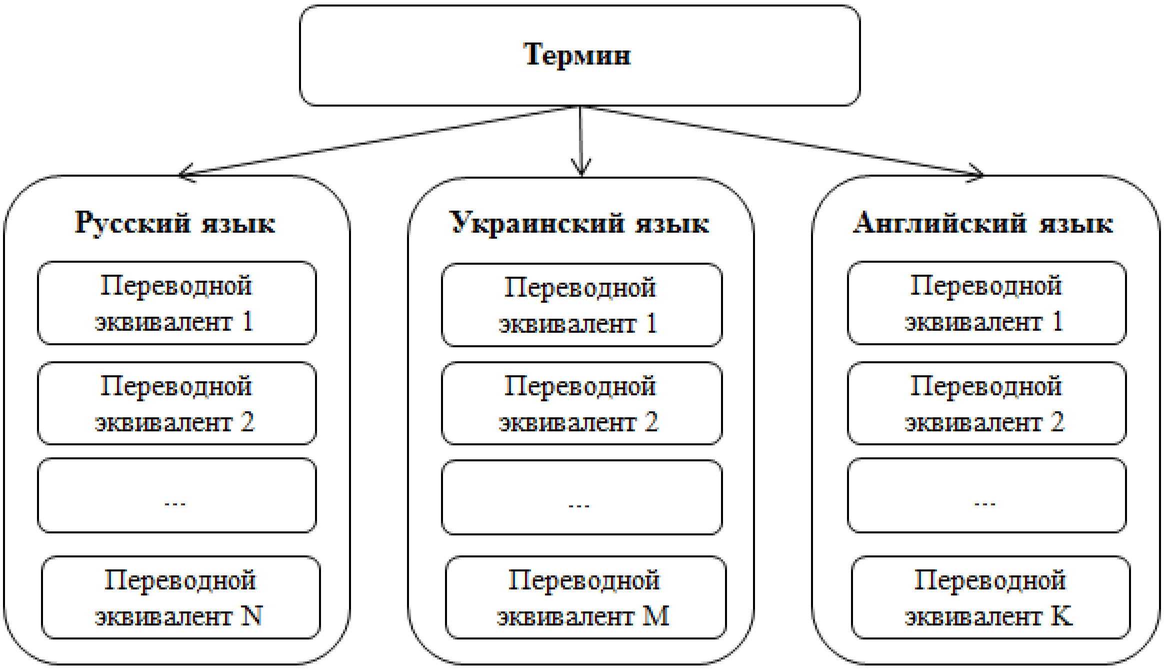


Рис. 3. Предлагаемая модель хранения данных

Таким образом, данный подход к построению трёхъязычного словаря позволит перейти от двуязычного словаря не только к трёхъязычному, но и многоязычному (с некоторыми ограничениями) словарю.

Заключение

Таким образом, были обработаны отсканированные тексты терминологического русско-украинского словаря по информатике и радиоэлектронике. Предложен подход к организации модели данных для трёхъязычного русско-украинско-английского терминологического словаря. При построении модели удалось уйти от некоторых проблем, которые возникали в ранее построенных словарях [1, 5]. Так, удалось автоматизировать корректировку входных данных [6], применив регулярные выражения в частности. Для формализации языковой информации, использовалось понятие семантического состояния языковой единицы. Очевидно, что поход к построению трёхъязычного словаря можно также применить и для словарей с большим количеством языков.

Достоинством данного подхода является равноправие языков, благодаря чему появляется возможность выбора основного языка, в зависимости от требований специалиста, работающего со словарем.

Литература

1. Широков В.А. Комп'ютерна лексикографія. – Київ: науково виробниче підприємство «Видавництво «Наукова думка» НАН України», 2011. – 352с.
2. Рабулець О.Г., Широков В.А., Якименко К.М.. Дієслово в лексикографічній системі – К.: Довіра, 2004. – 259 с.
3. Бондаренко М.Ф., Шабанов-Кушнаренко Ю.П. Теория интеллекта: учеб. – Харьков: Изд-во СМИТ, 2006. – 571с.
4. Бондаренко М.Ф., Шабанов-Кушнаренко Ю.П. Мозгоподобные структуры: справочное пособие. Том первый – К.: Наукова думка, 2011. – 460 с.
5. Остапова И.В. Лексикографическая структура этимологических словарей и их представление в цифровой среде // Прикладная лингвистика и лингвистические технологии: сборник научных трудов. – 2007. – С. 236-245.
6. Word VBA reference // [Електр. Ресурс]. – Режим доступу: https://msdn.microsoft.com/en-us/library/office/ee861527.aspx

Literatura

1. Shirokov V.A. Comp'uterna leksikographiya – Kyiv: naukovo vyrobnyche pidpryemstvo "Vydavnytstvo "Naukova dumka" NAN Ukrainy", 2011. – 352c.
2. Rabulets O.G., Shirokov V.A., Yakymenko K.M. Dieslovo v leksykographichniy systemi – K.: Dovira, 2004. – 259 c.
3. Bondarenko M.F., Shabanov-Kushnarenko Y.P. Teoriya intellekta: ucheb. – Kharkov: Izd-vo SMIT, 2006. – 571c.
4. Bondarenko M.F., Shabanov-Kushnarenko Y.P. Mozgopodobnye struktury: spravochnoe posobie. Tom pervyi – K.: Naukova dumka, 2011. – 460 c.
5. Ostapove I.V. Leksykographicheskaya struktura etimologicheskih slovarey i ih predstavlenie v cifrovoy srede // Prikladnaya lingvistika i lingvisticheskie tekhnologii: sbornik nauchnyh trudov. – 2007. – C. 236-245.
6. Word VBA reference // [Elektr. Resurs]. – Rezhym dostupu: https://msdn.microsoft.com/en-us/library/office/ee861527.aspx

RESUME

O.S Puzik, V.V Kurasova, G.G. Chetverikov

Aspects of data model organization of electronic trilingual dictionary

The article describes approaches to organize data model for trilingual dictionaries. The Russian-Ukrainian terminological dictionary of informatics and radio electronics in paper form is the base for electronic trilingual dictionary. The goal of the work is to extract data from scanned pages, perform corrections to eliminate or at least reduce number of incorrectly recognized words. Also the goal is to investigate aspects of data model organization for trilingual dictionary.

Stages of construction of electronic dictionaries are described as following:

a) Scanning and recognition of dictionary articles.

b) Correction of recognized texts.

c) Splitting texts to an array of separate article entries.

d) Decomposition of the array by formal characteristics.

Construction principle of the dictionary is alphabetic-nested. So heading term is Russian. The nest includes term's word-combinations which contain heading word as one of elements. Ukrainian translation equivalents correspond to order in Russian word combinations. Thus the construction principle is the reason why we can parse input automatically. Issues which prevent using data stored in "doc" format directly are analyzed. Office COM Automation technology is suggested as a way to handle scanned and recognized input data stored in "doc" format. Regular expressions are used as way to fix issues caused by recognition errors of OCR engines.

Two different approaches to store parsed dictionary data are considered. The article notes that increasing number of languages in multilingual dictionaries causes avalanche-like increasing number of relationships between translation equivalents. The study suggests extracting abstract semantic entity over each translation equivalent to avoid such issue. Words with the same semantic will construct the term entity. This approach allows expanding trilingual dictionary with other languages, explanations for each term and so on.

Надійшла до редакції 29.10.2016

